Middlesex University London

Middlesex University Research Repository

An open access repository of

Middlesex University research

http://eprints.mdx.ac.uk

Whitby, Blay (2003) The social implications of artificial intelligence. PhD thesis, Middlesex University. [Thesis]

This version is available at: https://eprints.mdx.ac.uk/7984/

Copyright:

Middlesex University Research Repository makes the University's research available electronically.

Copyright and moral rights to this work are retained by the author and/or other copyright owners unless otherwise stated. The work is supplied on the understanding that any use for commercial gain is strictly forbidden. A copy may be downloaded for personal, non-commercial, research or study without prior permission and without charge.

Works, including theses and research projects, may not be reproduced in any format or medium, or extensive quotations taken from them, or their content changed in any way, without first obtaining permission in writing from the copyright holder(s). They may not be sold or exploited commercially in any format or medium without the prior written permission of the copyright holder(s).

Full bibliographic details must be given when referring to, or quoting from full items including the author's name, the title of the work, publication details where relevant (place, publisher, date), pagination, and for theses or dissertations the awarding institution, the degree type awarded, and the date of the award.

If you believe that any material held in the repository infringes copyright law, please contact the Repository Team at Middlesex University via the following email address:

eprints@mdx.ac.uk

The item will be removed from the repository while any claim is being investigated.

See also repository copyright: re-use policy: http://eprints.mdx.ac.uk/policies.html#copy

MX 0106589 0

The Social Implications of Artificial Intelligence

Blay Whitby

The Social Implications of Artificial Intelligence

by

Blay Whitby

Context document, submitted in partial fulfilment for the degree

of

Doctor of Philosophy by publication

10/27/03

School of Computing Science
Middlesex University
London, UK

p1191456

Site	MIDDLESEX UNIVERSITY LIBRARY
Accession No.	U106589
Class No.	006.3 WHI
Special Collection	✓

X 303.4834

Acknowledgements

I must thank Dr. Carlisle George who has been the very exemplar of a supervisor; Professor Steve Torrance for constant input and encouragement; and Professor Maggie Boden for originally inspiring me to pursue this topic.

Contents

Site	NEW YORK UNIVERSITY LIBRARY
TM	
Accession No.	
Class No.	
Special Collection	

1. Introduction

For 18 years, I have been publishing books and papers on the subject of the social implications of Artificial Intelligence (AI). This is an area which is has been, and remains, in need of more academic attention of a serious nature than it currently receives.

It will be useful to attempt a working definition of the field of AI at this stage. There is a considerable amount of disagreement as to what does and does not constitute AI and this often has important consequences for discussions of the social implications of the field.

In brief, I define AI as the study of intelligent behaviour (in humans, animals, and machines) and the attempt to find ways in which such behaviour could be engineered in any type of artefact. My position on the definition of AI as a field of activity is set out in full in various places in the works submitted for this application. Most important are Chapter 2 of Whitby, 1988b, Chapter 3 of Whitby, 1996, and Whitby, 2000. This definition is distinctive (though not unique). For the purposes of discussion of social implications, its most distinctive feature is that it does not require the imitation or replication of *human* intellectual attributes. Because, under this definition, AI is not limited to the study of and attempt to build human-like intelligence the discussion of its social implications is rendered much broader.

Also, because AI encompasses the attempt to engineer intelligent behaviour in *any type* of artefact, discussion of its social implications will need to consider the way in which AI technology, methods, and attitudes can permeate other different areas. This will include a wide range of technologies which include an AI element and a wide range of disciplines which are influenced by AI ideas.

Thus the social implications of AI are turned into an immensely important field of study, since AI technology will steadily continue to permeate other technologies and thereby society as a whole. Many of the social implications of this technological process are non-obvious and surprising. If we are to make sensible, timely, and practical policy decisions and legislation then it is important to be as clear as possible about likely technological developments and their social implications.

We may initially attempt to characterise various approaches by other authorities on the social implications of AI. These range from the wildly speculative such as Warwick (1988) and Moravec (1988) to the mainly technical, for example Michie (1986). At the wildly speculative end of this continuum represented by Professor Warwick there are scare stories involving robots taking over the earth. (See for example Warwick, 1998 pp. 21-38) At the other end of the continuum, there are writers who often see AI as entirely positive, or as having no social implications at all.

Most authorities will, or at least should, occupy a position somewhere between these extremes. However, in giving serious academic consideration to this area, one needs to respond to this entire range of approaches. That is to say that one must (as a minimum) both be conversant with probable technical developments and also carefully and critically respond to speculations about the nature of future society. In my research I have consistently attempted to do just this.

This is obviously a cross-disciplinary exercise and the differing methodologies of different disciplines present further problems in determining the best (or an approximation to the best) approach. For a number of reasons, which will be fully explored in this statement, my approach has concentrated (mainly, though not exclusively) on the attempt to provide guidance to those actually concerned with the technical and scientific development of AI.

The published books and papers submitted as part of this application span a period of 16 years. These works form a coherent body of research around the area of the social implications of AI. This body both develops the theme of the need for professionalism in AI and answers the criticisms of other writers in the area. They involve a full response to other writers in this area, over the entire continuum described above. This is a large, coherent, important, and generally well-regarded body of work which is in every relevant sense equivalent to that required for a PhD. by research.

1.1 Overview of the document

The next section (Section 2) gives a background to and critical review of the works selected for this application.

Section 3, then establishes how the volume of work submitted for this application constitutes a coherent body of work.

Section 4 demonstrates and explains the way in which my work constitutes a significant and original contribution to knowledge.

In Section 5, evidence is provided that the portion of my research described in this document is equivalent to a PhD Thesis.

Section 6, discusses issues concerning my research methodology.

In Section 7 I provide a critical review of my personal development over the period spanning 16 years since the first publication submitted in this application.

Section 8, discusses some limitations of my work mainly related to interdisciplinarity and Technological change.

2. Background to and Critical Review of the Works Selected

In 1988 I published AI: A Handbook of Professionalism (Whitby, 1988b) which argued for the need for a more professional approach to the field by those actually working in it as scientists or technologists. The core argument and the code of conduct at the heart of this book had already been published in a peer-review journal (Whitby, 1988a) One principal claim of this book was that those working in AI should take responsibility for the wider social problems raised by their work. Many details of exactly how this should be done, including a proposed code of conduct for scientists and engineers working in AI were provided. This was a unique position among writers on AI at the time and remains a distinctive position. Although this book grew out of a large number of talks and papers I had previously produced on the subject of the social implications of AI, it was pivotal. The distinctive position of claiming that the social implications of AI should be managed initially by those actually working in the field has not been universally accepted

One challenge emerged initially in a review of Whitby, 1988b by Richard Forsyth (a lecturer at the University of the West of England and commentator on AI). Forsyth's view on the 'professionalization' of AI was that it would not be in the public interest. He saw professional bodies (such as the Law Society and the British Medical Association- BMA) as acting to further the interests of their members, usually to the disadvantage of the public. Aside from the implied criticism of the BMA and Law Society contained in such a comment, it does not attack the main thrust of my arguments. It had not actually been my claim that professional organizations such as the British Computer Society (BCS) will automatically bring about the improvements which I argued were necessary.

My response to this challenge has been to change the amount of emphasis which I give to codes and laws as opposed to high professional standards in general. In many ways it is better that people working within AI acquire a sense of social and moral responsibility about the consequences of their work. A code has a role to play in this, but some writers see codes as removing the need to think about the moral and social consequences of one's work.

This is a deep and complex issue which cannot be fully explored here. At the time I was preparing this book (Whitby, 1988b) I was directly involved in a movement within SSAISB (The Society for the Study of Artificial Intelligence and the Simulation of Behaviour). Some of us wanted to structure SSAISB more like a 'profession' such as law or medicine. The majority of committee members felt that SSAISB should remain more informal which it has indeed done. On the other hand, the BCS (British Computer Society) and IEEE (Institute of Electrical and Electronic Engineers) have taken a more profession-like approach. Both these bodies publish codes of conduct similar to the code proposed for AI in Whitby, 1988a. However, the BCS (correctly in my view) also requires the teaching of an element of professionalism as part of any degree that it accredits.

At this distance in time, my conclusion would be that the codes of conduct have a distinct part to play, but that they can never eliminate or replace the need for people to behave in a professional manner.

Another issue that was repeatedly raised in opposition to AI: A Handbook of professionalism (Whitby, 1988b) was that those working in the field of AI, either as scientists or engineers, had no real incentive to behave in a professional manner. The main motivation for innovation in the whole IT industry is commercial gain. Professional standards have, as a matter of history, been extremely low. This problem is still commonly cited whenever I give talks on this area and in written references to the works discussed above. This criticism is answered fully in two subsequent submitted works.

Firstly, the particular issue of the reasons for behaving in a professional manner are directly tackled in 'Ethical AI' (Whitby, 1991c). This paper argues that being ethical pays, both in business and academia. That is to say that it is in the long-term interests of those working in AI to behave in a much more professional manner. Business in general is moving towards being more ethically aware, and AI in particular would benefit greatly from doing the same. AI has the potential to be an extremely socially beneficial technology. There is much to be gained in commercial terms by presenting it to the general public as an ethical technology. Similarly, the academic AI community has much to gain from behaving in a more professional manner.

Secondly, in the first chapter of Whitby, 1996, a further response is made which is more specifically related to the technology of AI. The history of AI research, it is claimed, has seen too many exaggerated and over-optimistic claims about impending technological breakthroughs and the generalizability of current techniques. The inevitable disappointments and hostile reactions from funding bodies have meant that AI research has been subject to a stop-go funding pattern. Very often this has entailed researchers having to adopt a new 'fashion' in AI when previous methods had, rather too quickly, become discredited in the eyes of funding bodies.

Focussing on the behaviour of 'AI professionals' is only part of an analysis of the social implications of AI. In addition it is argued in the two books already cited (Whitby, 1988b, Whitby, 1996) that AI research takes place in a social context and is influenced in a number of ways by that social context. More importantly, for the focus of my research, AI has the power to influence that social context. AI has influenced the social context. The historical accounts of the importance of AI given in these two books give much emphasis to the way in which AI has been at least as successful in exporting ideas and methodologies to other areas as it has been in producing working technology.

Particularly relevant in this context are two papers on the complex relationship between AI and the legal system (Whitby, 1991a and Whitby, 1991b). These papers argue that AI will have a subtle and pervasive influence on legal practice, more through its influence as a group of new ways of looking at legal practice, than as a technology.

AI has two major legal effects. The first is as a technology that requires new legislation and the second is to provide new legal ideas. On the first point, there are several ways in which AI developments require consideration of legislation. In many ways AI requires the same legislative approach as Information Technology in general. However, it also has its own distinct legal implications. The most obvious of these is to prompt more consideration of the difficult issue of non-human agency. That is to say about items other than human beings acting as legal agents. This is not a totally novel legal notion but there are dangers in being too willing to accept a technological notion of non-human agency. The most important is the reduction of human responsibility, which will in turn affect legal liability. This interacts with further difficult technical questions about autonomy – how much autonomy does an AI program really have? There are also difficult human

questions to be faced in consideration of this issue. There is evidence examined in (Whitby, 1988b) that humans may tend to use any notion of non-human agency as a convenient screen behind which to hide their own culpability. Hasty decisions in this area could hide disreputable human motives in claiming that they exercised less responsibility than was actually the case.

This will require changes of attitude in both legal professionals and AI technologists. Some of the changes in attitude required of AI technologists had been fully described in Whitby, 1988b; important suggestions as to the changes in attitude required of legal professionals were made in Whitby, 1991a.

The importance of this particular area is emphasized by the approach of these papers (Whitby, 1991a Whitby, 1991b) to jurisprudential theory. It has been claimed (for example in Leith, 1986) that legal thought and practice is an area which is and should remain characterized by a process of social negotiation. It is not, therefore, suitable for simple rule-following AI approaches. AI practitioners need to appreciate this and their own role in the processes of negotiation.

The problems of introducing AI technology into the legal area are not only technical in nature. It is important to draw attention to questions of human accountability when introducing AI methods into the legal area. Questions about the appropriate degree of formalization or about the types of explanation which a system should generate are just as much legal and jurisprudential questions as they are technical questions. It is also important to have as wide as possible public debate on the issue of which parts of legal decision making are appropriate for the introduction of AI techniques and which are not. This, like so many others in the area of my research, is a fundamentally cross-disciplinary issue – a point which is further examined in section 6, Methodological Issues.

3. The coherence of the work

The works described in the preceding section form a single coherent body. They represent a particular approach to the social implications of Artificial Intelligence. In summary, this particular approach involves concentrating on the practical decisions immediately necessary to ameliorate the worst adverse social effects. This single brief account should not be taken as implying that there have not been significant developments in approach over the period of the work. There have been important personal and academic developments over the period which are fully discussed in section 7. There have also been significant technological advances in AI and related areas over this period and these have required attention in the context of the works under discussion.

The coherence of this entire body of work is demonstrated in various ways in this paper – particularly in section 6 which outlines a unifying methodology. However, the submitted works also pick up themes from earlier works. Some papers are direct responses to criticisms of earlier publications. This is particularly true of Ethical AI (Whitby, 1991c). This paper responds directly to the most frequent objection to previous works – namely that ethics count for little in the real world.

The overall coherence of the submitted works is best illustrated by consideration of Reflections on AI (Whitby, 1996). This book briefly outlines the distinct features of my methodological approach to this area and shows how my particular interpretation of the Turing test (included as Whitby, 2000, but first presented at the Turing Colloquium in 1990) is central to my entire approach to the social implications of AI. A framework is thus provided for discussion of the social implications of AI. This includes such issues as whether or not AI will have a different set of social implications from computing in general, whether or not AI systems should give advice in moral and emotional areas, and what sort of society will result from the widespread use of AI and related technology

4. Evidence and exemplification of claims made that the research constitutes a significant and original contribution to knowledge.

4.1 Summary of contributions to knowledge.

4.1.1 Focus on the social implications of AI rather than on IT or computing in general.

The distinctions are set out in full in Whitby 1988 Chapter 2 and Whitby 1996 Chapter 2.

4.1.2 Focus on the immediate practical consequences of AI.

This is in distinction to much philosophical discussion of technically remote or unlikely possibilities. A good illustration of this is the discussion of Whitby and Oliver 2000 in section 4.2

4.1.3 Exploration of the issues raised by non-human agency.

This is the main burden of Whitby 1991a and is explored in more detail in Whitby 1996.

4.1.4 Development of a novel methodological approach.

This is discussed and evidenced in detail in section 6 of this document.

4.1.5 Influence on the real world

This is evidenced by the media appearances detailed in appendix 2 and the contributions to teaching described in the following section (4.2).

4.2 Evidence and exemplification of contributions to knowledge.

My research, and particularly the papers selected from it for consideration here, represents a distinct and original contribution to knowledge. In brief this contribution is the focussing of philosophical debate on the immediate social and practical issues raised by the introduction of AI and closely related technology. A number of features distinguish my approach from other related investigations and debates. The most important are the concentration on immediate social consequences rather than distant or abstract possibilities and the focus on the role and responsibilities of scientists and technologists.

Also distinguishing this research are the specific questions about the particular social implication of AI rather than other science and technology. AI is distinct from other technologies. It deliberately sets out to examine the possibility of non-human agency. This is examined in detail in Whitby 1988b, Whitby 1991b, Whitby 1996, and Whitby and Oliver 2000. AI technology also has the power to reify and operationalize knowledge. This, in turn can give more legitimacy to those, such as legislators who wield power in the knowledge society. This special feature of AI and its effect on the balance of power in society are explored in Whitby 1991b and Whitby 1996.

The publications considered here also represent a distinct position on the social implications of AI. Though the position is arguably distinct, interest in this area is not unique to me. Other distinguished writers have published extensively in this and related areas. Notable among these are Maggie Boden, Dan Dennet, and Steve Torrance. There are significant overlaps in the approach of these three writers in particular and my approach. However, distinct to my approach is the emphasis on the need for researchers and engineers in AI to take most of the responsibility for the social consequences of the technology. Boden, Torrance, and Dennet write for a wider audience and place the burden of responsibility rather wider than my approach. That is not to say that I argue that wider society has no contribution to make to the debate. Far from it, general debate on these issues is to be encouraged and the feedback I have received indicates that my work has a wider audience than simply the professionals in the field. Other groups also have important responsibilities for the social implications of new technologies and this point is developed in section. 6

Particularly relevant to the context of general debate is the joint paper with Kane Oliver - a DPhil. student at Sussex University (Whitby, and Oliver, 2000). We wrote this paper as a response to the widespread attention given by the non-technical press to views about the social implications of AI which we considered unfounded. In particular, we were responding to Warwick, 1988 and Moravec, 1988. These writers have repeatedly attempted to scare the media and hence the public with claims that AI will shortly produce a situation where machines will dominate human beings. These claims are often considered so absurd as to merit no response from active AI scientists and technologists. Oliver and I felt there was a need to straddle the gap between the wider audience and the narrower technical audience with this paper. It says (primarily for the benefit of AI scientists and technologists) that these claims are not absurd; they are merely false and therefore should be actively denied. It says, for a wider audience, that there is no cause for concern or legal and political control of AI on the basis of these false claims.

Neither Warwick nor Moravec has felt the need to respond directly to our paper but it is referenced on the web-site of Professor Warwick's research department. It has also been responded to by Henry Cribbs (Cribbs, 2000) Cribbs cites our paper as "doing a good job of defusing" the arguments in favour of a possible robot takeover. Out paper is also included in the 'Robots won't rule' web site maintained by Chris Malcolm at the University of Edinburgh.

My research has also had some influence on the real world. In particular I am often invited to comment on the social implications of AI and on Information Technology in general by journalists. A list of TV and radio appearances specifically as an expert in this area is attached as Appendix 2. In addition it is clear that courses on the professional aspects of computing and IT have become far more common as part of undergraduate degrees in computer science. I cannot claim sole credit for this, but I have been an enthusiastic participant and have delivered such courses for twelve years first at Middlesex University and lately at Sussex University. Obviously, my research feeds into these courses and I can at least hope that the many successful graduates of my courses now working in the IT industry have been influenced to some degree.

5. Evidence that the research described is equivalent to a thesis.

The submitted works are a large and important body of work which is in all senses equivalent to a thesis. The word length of these publications exceeds that which would be required for a thesis. Amusingly in this context, one reviewer thought 'AI: A Handbook of Professionalism' (Whitby, 1988) was, in fact, a book made from a thesis. This work has the general structure of a thesis in that it critically reviews relevant literature, has a distinct methodology, and performs an analysis using this methodology in order to reach a clear set of conclusions. It also makes a novel and original contribution to knowledge, as has been illustrated in the preceding sections.

The papers submitted have been subject to peer review in that they have been published in reputable conferences or journals. This in many ways parallels a thesis examination. The collective number of issues discussed in the submitted works is representative of the sort of coverage required for a thesis.

The focus of the work as described in section 1 is appropriate for a thesis. This is most clearly seen in consideration of Reflections on AI (Whitby, 1996). This book gives a full and coherent account of the views set out in papers published over the period 1989-1996. The research was carried out over this period. It is the primary publication on which the application is based. Most of the claims made in this book are strongly founded on those made in peer-reviewed conferences and journals. My analysis of the Turing test (Whitby, 1996, Chapter 3) was considered worthy to be included in a collection of papers which have shaped the history of cognitive science (see Whitby, 2000). This is some measure of its seriousness and value within the academic community.6. Methodological Issues

The methodologies used in my research stem from a foundational belief that technology should be created and used *only* for the benefit of humanity. Since this is a foundational belief I can only assert it here and offer no justification of it. However, as foundation, it shapes the methodologies actually chosen in the research.

The core methodology is a process of reasoned argument. In essence this is the technique of scholarly disputation most typical of philosophy as an academic discipline. In as much

as my research can be classified, the best (or least inaccurate) description would be *philosophy of technology*.

This foundational belief and core approach shape the research in many important ways. Firstly there is an attempt to focus on issues of practical social relevance. Many of the debates that interest philosophers of science and technology do not seem to be of practical social relevance. For example philosophical debates on the possibility of conscious machines are not likely to have social consequences for the foreseeable future. The research described in this document, by contrast, has attempted to focus on aspects of technology that have more immediate social consequences. This core belief and distinctive approach have two further methodological consequences – the importance of human choices and the need for eclecticism which are discussed in detail in the following sections.

This eclecticism is, in many ways, distinctive and appropriate for this particular area. It not only recognizes the need for a genuinely interdisciplinary approach, but also pursues the subject matter over disciplinary boundaries in spite of the difficulties raised. The area of the social implications of technology requires (at least) both technological and social analysis. For this reason, among others, my work crosses a number of disciplines with differing, if not conflicting methodologies. It would be very difficult, and inappropriate, for me to adopt only the methodology of the natural sciences, or of the social sciences, or of philosophy. Some sort of mixture of methodologies is often called for in this area.

However, eclecticism is not, in itself, a methodology. It would be better described as an approach required as a consequence of the foundational belief and core approach outlined above.

An illustration of this is Whitby, 1993. This paper has a narrower focus, in some senses, than the other papers submitted here. It sets out to analyse the ethical implications of Virtual Reality (VR). Despite this narrower focus, it is necessary both to outline the present and likely future technological development of VR and to consider a number of theories of meta-ethics which might bear upon an ethical consideration of these developments.

The case of two disciplines is relatively easy to appreciate. The preface to Whitby, 1996 mentions the need to cover a number of specific disciplines, though this is not an attempt to produce an exhaustive list. This preface, and indeed the whole book (Whitby, 1996) suggest rather strongly that a single-discipline approach to this area would be ineffective, if not absurd. In short, a distinctive feature of my methodology has been to freely ignore the supposed boundaries between disciplines. This cross-disciplinary approach is distinctive and important to my research. There are, of course, various problems raised by an interdisciplinary approach which are discussed in section 8.

The second important feature which distinguishes my approach is its focus on the importance of those actually working in the area. That is to say that my research focuses on the professional responsibilities of researchers and technologists, rather than on advice to governments and lawyers.

This might seem to be in tension with the importance of human choices which was stressed in the penultimate paragraph. In fact there is no real tension here. The claim is merely that many important human choices in shaping the future of AI technology are those made by the people actually working closely with the technology. This is, at least partly, because their choices are technologically informed. In such a fast-moving and often technically difficult area it has been relatively difficult for those without the relevant technical knowledge to comment in detail. It is also important to note that, especially in the case of AI, technical decisions can have social consequences. Systems often incorporate many social assumptions at the design stage. Certain types of AI systems have the ability to shift the balance of power between different social groups. This point is explored in detail in Whitby 1991a, Whitby 1991b and Whitby 1996.

Although governments have often demonstrated a relatively poor historical performance in shaping previous technological developments, they have an important role in this area. This includes providing an appropriate legislative framework, assessing overall risks and benefits, and assuming the ultimate responsibility for social decisions. The responsibilities of the technologists - which are the focus of my research -include taking a professional attitude to their work and keeping governments, courts, and the public technically informed. . Governments and courts decisions can only be useful in

proportion to the accuracy of the technical advice received..

Despite my personal focus on the role of technologists, it also the case that various other groups can take decisions which affect the future of society and even the future of technology. These groups include the key players in authority (governments, legislators and the judiciary) and also many commercial and military interests. One of the ways in which can happen include the deliberate setting of technological goals by governments.

One of the most spectacular recent examples of this is President J.F. Kennedy' declaration on May 25th 1961of his county's commitment to a manned lunar mission. In this case there is clear evidence that a particular political decision drastically forced the pace of scientific and technological developments. In less spectacular cases the influence of political decisions may be harder to see but it is just as real. Some research areas may be neglected because of a lack of political or commercial interest. On the other hand, as with the US Apollo moon missions, an area may find political favour and the pace of development be much accelerated.

6.1 Reasons for choosing this methodology

The best way of explaining the advantages of my methodology is by examination of the problems of other methodologies commonly employed in this area. As was said in the introduction, other writers have approached this area in a number of differing ways. Some may adopt the approach of looking at likely technological developments and extrapolating these into the near and mid future. This approach has often been taken by those close to the field, for example Michie, 1986 and Partridge, 1986.

It would appear, prima facie, that their approach has several advantages. In particular, it is more likely to be technically informed. That is to say that it reduces the tendency to discuss purely 'science fiction' possibilities, since the technical predictions of those close to the area are likely to be more accurate. There should be a tendency to avoid wild speculation about social problems which might be raised by technology which is not in

fact in prospect. This, unfortunately, has not always been the case in practice. An illustrative example of a falling away from this principle is provided by The Creative Computer (Michie and Johnson, 1984).

A further methodological difficulty in this arc is that many people in AI have succumbed to the temptation to discuss exciting future possibilities of the technology, rather than concentrating upon more mundane immediate social implications. Examples abound, but Warwick, 1988 and Moravec, 1988 are the most relevant to this document. This is a recurring and central problem for serious discussion of the social implications of AI which I have always tried assiduously to avoid. These writers, of course, are contradicting my principle of concentrating first on problems of immediate social relevance. This is why some of the works included in this submission (Whitby 1996, Whitby and Oliver 2000) are direct attacks on them

A further problem with the mainly technological focus of this methodology is that it tends to attribute a very small role to human decisions about the future shape of society, and just as importantly, about the future progress of technology. Of course, what turns out to be *technologically* feasible is not primarily determined by human decision-making. To those close to the technology this is often the most important consideration. They tend to see progress as a mainly technological process. Progress is, at least partly, constrained by what is technologically possible.

This is particularly noticeable in the case of AI. It has become a truism of AI that certain technological achievements turned out be easy whereas certain others, predicted to be easy, have turned out to be extremely difficult. Examples are many, but one is the apparently insurmountable difficulty of enabling computers to understand natural language. Close to forty years of research have not produced computers that have anything approaching the facility of even a young child with human language. For the purposes of the present argument, writers who based predictions of the social implications of AI upon the achievement of this technological goal would now appear misguided and irrelevant. The need to be technologically realistic is also important in this area.

Some other important observations may also be made about this example. First the fact that it is an extremely difficult technological goal does not imply that it is impossible. In fact slow (very slow) progress has been made in the field of NLP (natural language processing). There is no prima facie reason to believe that it is impossible, rather than extremely difficult.

Second, the obvious ease with which human beings approach the task of understanding human language may be highly misleading. Humans almost invariably learn to speak at least one natural language before learning to play chess. A focus on human ability as the model for AI, would therefore suggest that language was the easier of the two tasks. In fact AI has found chess-playing the easier and it is good example of an area where AI can be taken to have been completely successful. It is important not to take too anthropomorphic a view of likely technological progress. One of the main causes of the overly human-centered view of AI is the influence of a paper by Alan Turing (Turing, 1950) and the so-called 'Turing test' which is why the analysis of Turing's 1950 paper in Whitby, 2000 is central to my research and underpins the attack on competing methodologies in this section.

The influence of this paper on both AI technologists and on commentators from other disciplines has been so profound that it is worth examining in some detail. In the paper (Turing, 1950) Turing says he wishes to consider the question 'Can machines think?' but considers this question too vague to be answerable. He therefore proposes replacing the question with a game (The Imitation Game) which involves a conversation with an unseen correspondent. What would we conclude, Turing asks, if an average interrogator could not reliably distinguish a human from a computer in this game?

It is easy to misinterpret Turing's paper. The most important way, for present purposes, in which this paper has been misinterpreted, is that of suggesting that the ultimate goal of AI is an imitation of human intelligence. The adverse consequences of such interpretations are discussed at length in Whitby 2000 (submitted as part of this application). This paper argues that AI has been mistakenly too focussed on the imitation of human intelligence. To return to the present example it can be seen that the focus on human imitation may have misled AI technologists into believing that natural language processing was much easier than it, in fact, turned out to be.

The fascination inside AI research with human-like intelligence in artefacts, rather than truly 'artificial intelligence' has a distorting effect upon predictions of the social implications of AI. This distortion involves a tendency to discuss the social implications the arrival of substantially human-like artefacts. Extreme examples of this are provided by the 'robot takeover scare stories' such as Warwick 1988.

Commentators close to the technology mentioned above are not immune from this distortion, though perhaps less susceptible. It is important to stress that most of the real social implications of AI are not associated with human-like artefacts. The achievements of AI are not obviously human-like. Indeed, the lack of superficial resemblance between human intelligence and artificial intelligence has led to the creation of the myth that 'AI has failed'.

Both the failure myth and the robot takeover myth need to be avoided completely in consideration of the social implications of AI. I have tried to debunk these myths and the misunderstandings which have caused them in Whitby, 1996, Whitby, 2000, and Whitby and Oliver, 2000. The real achievements of AI have important social implications and distracting myths must be avoided.

7. Critical review of personal development over the period.

The fifteen years over which the publication submitted have been produced is certainly long enough for research to develop changes of perspective. Although I have remained constant in certain core values – notably that vested interests and the technology itself should be subservient to the needs of humanity as a whole – there have been developments in my approach.

These developments can, perhaps, best be characterized by dividing the research into four main time periods. It is difficult to put exact dates on these periods as, of course, changes were progressive and there was a certain amount of overlap. Nonetheless, I believe that they form a good basis for an analysis of the development of the research

over the period of the submission.

7.1 The Four Main periods

7.1.1 The Outset (1984-1990)

During this period I wrote primarily as a social philosopher. Some of the carliest papers werc written from a slightly outsidcr perspectivc. That is not to say that that I was an unfriendly commentator (as were many philosophcrs at the timc). These papers (can you cite them here?) treat AI as a worthwhile enterprise and capable of achieving success, but draw attention to some important social consequences. These included the implicit elitism of those scientists at the cutting edge of AI research and the significance of the high proportion of militarily fundcd research in AI.

7.1.2 Practical Applications (1990 -1994)

During thc second period I acquired more knowledge of and skill in actually building AI programs. This led to interest in practical applications of AI. In 1991, for example, I developed an AI program which simulated some aspects of moral reasoning. This work is described in Reflections on AI (Whitby, 1996). The practical applications of AI in the moral and legal areas do not form part of the focus of this document, but are nonethelcss important in describing my development as a rcsearcher. This technical background meant that there was a progressive change of perspective. No longer did the publications adopt the stance of an 'outsider' critiquing AI, but acquired more of an 'AI insider's perspective. This was particularly noticeable at interdisciplinary conferences, where I would often be seen more as an apologist for AI than I had been previously.

7.1.3 An AI Insider (1994-2000)

During this later period I tended to be seen much more as someone representing the AI community. During this period I havc held a lectureship in AI at a prominent research centre. From 1996 I have been editor of AISBQ (The Quarterly journal of the Society for the study of AI and the Simulation of Behaviour). This is essentially the 'trade journal' of the UK Al research community and the post of editor definitely qualifies me as an 'insider'. This has tended to make me much better qualified to comment on the methodological direction of AI.

Situated squarely within this period is 'Reflections on AI' (Whitby, 1996). This book combines both the social comment of the earlier work and the discussion of AI methodology and philosophy from more of an insider's perspective. The expression 'insider's perspective should not be read as implying uncritical. Many portions of Reflections on AI (Whitby, 1996) were highly critical of basic assumptions of current AI research.

7.1.4 Synthesis (2000- 2002)

The fourth period is essentially an attempt at a synthesis of the above perspectives. My current research aims to combine the social philosopher's perspective on AI together with a technical appreciation of the state of the art which comes from being (at least partly) an insider. An example is provided by the latest paper submitted a part of this application (Whitby and Oliver, 2000). This paper deals wit h issues which are undoubtedly the province of social and political philosophy. However it is, at the same time, very much a statement by an 'AI insider' in response to predictions made by other AI insiders on the social consequences of AI research.

I am also writing, under contract, a general audience book - AI: A beginner's Guide (not submitted as part of this application). This is a task for which I now feel well qualified, and which demonstrates the sort of synthesis which characterizes this fourth phase of my development as a researcher. My research and publication continues to develop along these lines.

8. Limitations

8.1 Interdisciplinarity

A major limitation of this work stems from its intrinsically interdisciplinary nature. Viewed from a particular disciplinary specialism it will very often appear not to do justice to that specialism's methodological investment in the subject matter. For example, a purely legal approach to this area would probably start by reviewing current legislation incorporate the necessary amount of jurisprudential theory with perhaps a nod towards political considerations. This would be valid and interesting, but it represent only small fraction of what I am trying to do in my research.

Inevitably disciplinary boundaries will be precious to some commentators. In addition, truly interdisciplinary peer-review journals are surprisingly rare. The nature of my methodological approach makes research in this area much more difficult than it would be if placed centrally within a single discipline. However, given the importance of the area and the appropriateness of the methodology employed, it seems a price worth paying.

8.2 Technological change

The preceding sections have tended to treat AI as a fairly static entity. This, of course, is far from the case. The scientific and engineering activities that constitute AI have changed a good deal since 1980. In particular, new sets of approaches, such as Alife, situated robotics, and autonomous agents, have come into existence. This issue is specifically addressed in Chapter 2 of Whitby, 1996. It is an issue that requires continuing attention.

Consideration of the social implications of these new approaches reveals many similarities to the issues considered above. However, there are also many fundamentally novel moral, legal, and social questions raised by these new approaches. In a forthcoming paper I hope fully to address the novel problems raised by Alife and situated agents. This means that the bulk of this paper would represent essentially bringing the previous works up-to-date. I would expect this to be of sufficient quality to merit publication in a quality peer-review journal. As a subject area, it is both topical and in need of further serious

study. As editor of the UK's longest-standing AI journal, I remain in touch with current developments both in the UK, and overseas. I hope to continue research and publication in this area.

References

Cribbs, H., 2000, Bearing Frankenstein's Children: Artificial Intelligence and Universal Moral Values. *Ethics of AI discussion forum* (http://www.justlikethat.com/client/philosophy_user/ai.html)

Moravec H., 1988, *Mind Children, The future of robot and human intelligence*, Harvard University Press, Cambridge Mass.

Michie, D and Johnston R, 1984, *The creative computer: machine intelligence and human knowledge*, Viking, London

Michie, D. 1986, *On Machine Intelligence*, Ellis Horwood, Chichester.

Partridge, D., 1986, *Artificial Intelligence, applications in the future of software engineering*. Ellis Horwood, Chichester.

Turing, A, 1950, Computing Machinery and Intelligence, *Mind* Vol LIX. No. 236

Warwick, K., 1998, *In the Mind of the Machine*, Arrow Books, London.

Whitby, B. 1988, *AI: A handbook of professionalism*, Ellis Horwood, Chichester

Whitby, B. 1991a, AI and the Law: learning to speak each other's language. In A. Narayanan (Ed.), *Law, Computer Science and Artificial Intelligence*, Volume I. New Jersey: Ablex Publishing Corporation.

Whitby, B.R. 1991b, AI and the Law: Proceed With Caution. In M. Bennun (Ed.), *Law, Computer Science and Artificial Intelligence*, Volume II. New Jersey: Ablex Publishing Corporation.

Whitby, B. 1991c, Ethical AI, *Artificial Intelligence Review*, Vol.5, No.1.

Whitby, B. 1996, *Reflections on AI, The legal social and moral dimensions*, Intellect Books, Exeter.

Whitby, B.R. 2000, The Turing test: AI's biggest blind alley? In Chrisley R. (ed) *Artificial Intelligence: Critical Concepts in Cognitive Science*, Routledge, London pp. 195-212

Appendix 1

List of works submitted as part of this application

Whitby, B.R. (2000), The Turing test: AI's biggest blind alley? In Chrisley R. (ed) *Artificial Intelligence: Critical Concepts in Cognitive Science*, Routledge, London pp. 195-212

Whitby, B.R. and Oliver K. (2000) 'How to Avoid a Robot Takeover: Political and Ethical Choices in the Design and Introduction of Intelligent Artifacts' *Quarterly Journal of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour*, 104, Summer/Autumn 2000.

Whitby, B.R. (1996), *Reflections on Artificial Intelligence: The Social, Legal, and Moral Dimensions*. Oxford: Intellect Books.

Whitby, B.R. (1993). The Virtual Sky is not the Limit - The Ethical Implications of Virtual Reality. *Intelligent Tutoring Media*, Vol.3 No.2.

Whitby, B.R. (1991c). Ethical AI, *Artificial Intelligence Review*, Vol.5, No.1.

Whitby, B.R. (1991b). AI and the Law: Proceed With Caution. In M. Bennun (Ed.), *Law, Computer Science and Artificial Intelligence*, Volume II. New Jersey: Ablex Publishing Corporation.

Whitby, B.R. (1991a). AI and the law: learning to speak each other's language. In A. Narayanan (Ed.), *Law, Computer Science and Artificial Intelligence*, Volume 1. New Jersey: Ablex Publishing Corporation.

Whitby, B.R. (1988b). *AI: A Handbook of Professionalism*, Ellis Horwood, Chichester

Whitby, B.R. (1988a). A code of professionalism for AI. *Quarterly Journal of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour*, No. 64, Spring 1988, pp 9-10.

Whitby, B.R. (1987). Professionalism and AI. *Artificial Intelligence Review*, Vol.2, No.2, pp 133-139.

Appendix 2

TV and Radio appearances as expert on the social implications of AI and IT

Afternoon Shift, BBC Radio 4, Discussion on the importance of the Turing test, broadcast 17.1.97

Science Now, BBC Radio 4, Interview on the significance of the Turing test, broadcast 7.3.98

Tommorow's World, BBC1, Interview on the significance of the Turing test, broadcast 11.3.98

Connections, BBC Radio 4, Lengthy interview on the Social Implications of AI, broadcast 7.5.98

Working in IT: The Future, BBC Education, Lengthy interview on possible future developments in Information Technology and their social implications, broadcast 09.11.00

Discovery Today, ITN Factual, Panel discussion on the future development of Information Technology, recorded 19.12.2000

The Battle of the Robots – The Hunt for AI Channel 4 broadcast 13.10.2001

Open University TV series on the Future of IT recorded 4.2.2002